# Recommendation of Academic Collaborators: A Methodology Incorporating Word Embedding and Network Embedding

Xiaowen Xi[1], Ying Guo[2] and Weiyu Duan[2]

[1] Archives of Chinese Academy of Sciences, Beijing, China
[2] China University of Political Science and Law, Beijing, China
guoying_cupl@126.com

**Abstract:** Fruitful academic collaborations have become increasingly more important for solving scientific problems, participating in research projects, and improving productivity. As such, frameworks for recommending suitable collaborators are attracting extensive attention from scholars. In an effort to improve on the current solutions, we developed an approach that produces recommendations with better precision, recall, and accuracy. Our strategy is to leverage the benefits of the two most common similarity indicators for collaborator recommendation – research interests and co-authorship network topology into one unified framework. A Word2Vec model creates word embeddings of research interests, which solves the problem of calculating similarity solely based on co-occurrence, not context, while a Node2Vec model automatically extracts and learns the topological features of a co-authorship network, moving beyond just local features to capture global network features as well. The two similarity measures are then fused with CombMNZ resulting in a ranked list of recommended collaborators for the target scholar(s). The workings of the framework and its benefits are demonstrated through a case study on academics in the field of intelligent driving and a comparison with the two most commonly used baselines: Random Walk with Restart (RWR) and Latent Dirichlet Allocation (LDA). Our framework should be of benefit for academics, research centers, and private-enterprise R&D managers to find partners, so as to achieve the ultimate goal of completing research projects, solving scientific problems, and promoting discipline development and progress.

**Keywords:** Academic Collaborator Recommendation; Research Interest; Network Topology; Word Embedding; Network Embedding.

## 1 INTRODUCTION

The complexity and diversity of academic activities are ever-expanding, yet, with each new breakthrough, the intersections between disciplines are becoming more and more obvious. While not without its advantages, the increasing level of crossover necessary to find comprehensive solutions is making scientific research more difficult and

often beyond the capabilities of a single researcher or a research institution[1, 2]. Hence, academic cooperation has gradually become the modus operandi for conducting research. As Guns and Rousseau state[3], academic cooperation helps to improve the efficiency of scientific inquiry and research output[4]. In this vein, scholars and scientists, just like the professionals in any sector, typically aspire to collaborate with the highest-level researchers in their field possible. Often, the aim is to establish a joint research team to exchange knowledge, share resources and, hopefully, use the power of new and different perspectives to generate thinking greater than the sum of its parts. The ultimate goal, of course, is to successfully complete research projects, find high-quality solutions to scientific problems with greater efficiency, and to contribute to the development and progress of the entire field.

However, accomplishing all these objectives depends on identifying advantageous collaborators in the first place. Thus, recommendation frameworks for screening potential scientific collaborators have been a topic of intense focus for some time. Existing systems generally fall into one of two categories: those that recommend collaborators based on similar research interests[5, 6] and those that explore co-authorship networks[7, 8]. Frameworks based on similar research interests are typically built around text mining techniques that extract topics via keywords, subject terms, or labels. Recommendations are then calculated based on co-occurrence indexes or the like. The problem is that these types of indicators cannot capture the context in which the topic was mentioned, and so cannot factor that information into a recommendation. This can easily result in an inaccurate representation of a scholar's research interests and, thus, inappropriate recommendations. The frameworks designed to explore co-authorship networks generally base their similarity calculations on the network's topological features, using indicators like the Common Neighbor Index (CN), the Restart Random Walk Index (RWR), the Local Path Index (LP), and so on. The problem with these approaches is that, first, each recommendation problem requires its own custom-refined indicator or set of indicators, and designing the 'perfect' indicator is a task that requires a great deal of finesse. Second, current topological indicators only capture local features, such as direct paths or common neighbors or others. They do not typically provide "big picture" information about the entire network.

With the aim of addressing these concerns, we developed a novel framework for recommending academic collaborators that leverages both types of indicators through word and network embedding. There are three main steps to the process: 1) extract the research interests of scholars from a corpus of articles with the Word2Vec model, then calculate the similarity between scholars' interests in terms of cosine distance; 2) construct a co-authorship network, then extract and calculate the similarities between topological features with the Node2Vec model; and 3) integrate the results of both similarity measures using the CombMNZ method to produce a ranked list of recommendations for the target scholar. Additionally, to verify the effectiveness and efficiency of our framework, we conducted an empirical analysis on the field of intelligent driving and compared the results to the two most common recommendation approaches for finding potential collaborators used today. The results show our recommendations have higher accuracy, recall rate and F1. The approach can be applied to a range of fields/sectors/industries with little to no modifications. Academics, research centers, and private-

enterprise R&D managers should find the insights and recommendations provided by our system highly useful.

## 2 RELATED WORKS

In early studies on general recommender systems, the scholar's research interest was mainly captured by extracting salient terms and phrases from the dataset. The next main advancement came with feature weighting through techniques such as Term Frequency-Inverse Document Frequency (TF-IDF)[9]. However, due to the ambiguity of natural language (synonyms, polysemy, etc.), comparing scholars based on keywords does not always accurately reflect the actual similarity of their academic interests. Topic models are credited with solving some of these ambiguity problems with language and also for raising general interest in feature extraction[10]. However, topic models treat documents as a bag-of-words and assume that words occur independently. Without context, the recommendations produced cannot be completely reliable. The Word2Vec model, however, is an efficient word embedding technique that is able to learn the semantics of terms in context and form a dense, low-dimensional vector for each word[11]. Thus, we apply Word2Vec model to extract more finely-grained features representing the research interests of scholars in our framework, which substantially improves the accuracy of the recommendations.

Among the network analysis approaches to collaborator recommendation, co-authorship prediction is an important line of work. These studies have employed and combined several similarity indicators in co-authorship network to predict and recommend collaborators[12]. For example, kong et al. used Random walk with restart model (RWR) to measure the academic impact of researchers on the collaborator network[13]. However, authors' features mainly depend on manual design and selection, so it is necessary to realize automatic extraction of network topology features. In addition, the computational complexity that comes with increasing dataset remains a complicated and difficult task. The Node2vec model could transform the semantic information of nodes in the original network into a low dimensional vector space and effectively preserves the network structure of nodes, which can efficiently calculate the semantic connections between nodes in the network[14]. Thus, to extract features automatically and accurately, the Node2vec model is exploited to generate feature vectors of scholars' network topology.

## 3 METHODOLOGY

The recommendation framework for academic collaborators based on word embedding and network embedding model proposed in this paper comprises four main steps: data acquisition and preprocessing; measuring the similarity of research interests between scholars; measuring the similarity of topological features between scholars; and recommending suitable academic collaborators with the CombMNZ model. An overview of the framework is provided in Figure 1.

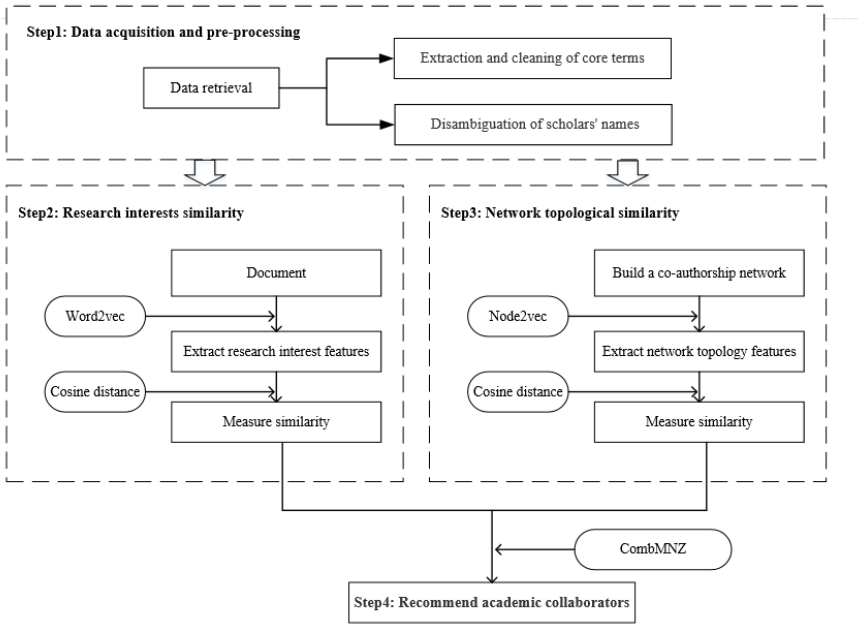


**Fig. 1.** Analytical framework

### 3.1 Data acquisition and preprocessing

This step involves retrieving and downloading academic articles from the Web of Science database, then using a professional desktop text mining software—VantagePoint (https://www.thevantagepoint.com/)—to extract key features such as the author, year of publication, title, and abstract. With a raw dataset of terms assembled, the subsequent data preprocessing procedure cleans the terms and disambiguates the author names in two separate steps.

The procedure of cleaning terms is as follows. First, the title and abstract fields are merged, and VantagePoint performs word segmentation. Noise is then removed and synonyms are merged with a term clumping process based on a fuzzy semantic matching algorithm developed by Zhang et al.[15]. Terms and phrases appearing more than six times are then extracted for further analysis by experts who remove general and irrelevant terms, such as development, methods, significant, etc. The final culled list forms the vocabulary of core terms.

Authors with the same names are disambiguated through a two-dimensional matrix where the rows contain the names and the columns contain their affiliations. A fuzzy matching algorithm then merges duplicate scholar names and institutions.

### 3.2 Research interest similarity

Word2Vec focuses on sequential combinations of words in a corpus and exploits the idea of neural networks to train a language model that maps each word to a vector.

Word2Vec includes two model options for updating parameters to suit different situations[16]. For our purposes, the training procedure in Skip-gram produces a more accurate result.

To obtain accurate eigenvectors of the scholars' research interests, this step is to produce accurate eigenvectors of scholar's research interest. More specifically, given a series of documents D $= \{d_1, d_2, \ldots, d_n\}$ with a vocabulary of N words$\{w_1, w_2, \ldots, w_n\}$, the Word2Vec model maps each word in the vocabulary to a fixed-length vector $\{v(w_1), v(w_2), \ldots, v(w_n)\}$ based on the co-occurrence relationship between documents and words. The document vector $\mathrm{v}(d_i)$ is then calculated by plusing each word vector as follows:

$$v(d_i) = \sum_{n=1}^{m} v(w_n) \tag{1}$$

Where $m$ represents the number of words in the document.

The author vector $\mathrm{v}(c_i)$ is then computed by plusing each document vector according to the co-occurrence relationships between documents and authors as follows:

$$v(c_i) = \sum_{i=1}^{n} v(d_i) \tag{2}$$

Where $n$ is the number of documents written by the author.

With the fixed-dimension feature vectors of the research interests generated, the next step is to calculate the similarity of interests between researchers. Of the many methods of measuring similarity, we chose the popular and widely-used cosine similarity index, formulated as

$$sim(A,B) = cos(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{j=1}^{m} a_j b_j}{\sqrt{\sum_{j=1}^{m} {a_j}^2} * \sqrt{\sum_{j=1}^{m} {b_j}^2}} \tag{3}$$

Where the vector of author A is ($a_1$, $a_2$, …, $a_m$), and the vector of author B is ($b_1$, $b_2$, …, $b_m$).

### 3.3 Topological similarity

Node2Vec has been proven to maximize the likelihood of preserving network neighborhoods and also can maps the nodes to a low-dimensional feature space[14]. To efficiently and effectively acquire feature of scholars' network topology, we take a co-authorship network $G = (V, E)$ as input, after running the node2vec model, we can get the $n * \beta \mathrm{matrix} N_{raw}$ to represent his/her network topology feature, where $n$ regards the number of nodes, $\beta$ is the parameter that determines the dimension of the node's vector representation, and the final output is $N_{raw} = \{v_1, v_2, \ldots, v_n\}^T$.

Calculating the cosine similarity between the topologic features of each scholar follows the same basic principles as with the research interests described in Eq. (3).

### 3.4 Recommendations with CombMNZ

The goal in this stage is to integrate the two similarities and rank the candidate collab orators from high to low according to their similarity. We have opted for a score-based

algorithm because it is the most widely used in the field of recommendation and, more specifically, CombMNZ[17].

To fuse the similarity results with CombMNZ in a fair way, the dimensions of each similarity measure first need to be standardized. The CombMNZ calculation is then[18]:

$$score_{combmnz} = n(j,w) * \sum_{n=1}^{N} m_n * score_{normal}(j,w_n) \quad (4)$$

where $n(j,w)$ denotes the number of times scholar $j$ appears in the score $w$ of each dimension, $score_{normal}(j,w_n)$ denotes the standardized score of the scholar in the $w_n$ item ($w_n \leq 2$), and $m_n$ denotes the weight of each dimension derived with a greedy strategy.

# 4 CASE STUDY

## 4.1 Data collection and preprocessing

To assemble our corpus, we retrieved papers published between 2010 and 2018 from the Web of Science database using a search strategy drawn from Kwon et al. as follows[19]:

TS=(((Self-driving or autonomous or driverless) near/4 (transport* or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane))) or TS = (((drone near/2 autonomous) or (uav near/4 autonomous))) or TS = ((robot* near/1 (transport* or mobile or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane)) AND (autonomous or self-driving or driverless)) or TS = (“autonomous driv*”) or TS = (((robot* near/1 (transport* or mobile or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane)) OR (drone or uav)) AND (path or planning or planner or plan)) or TS = (((robot* near/1 (transport* or mobile or car or motorcar or vehicle or automobile or aircraft or airplane or aeroplane)) OR (drone or uav)) AND (2D or 2-D or 3D or 3-D or map or localization or tracking or navigat* or obstacle or avoid*)).

The search returned 34,244 records. NLP preprocessing with VantagePoint yielded 8,637 core terms and phrases. Outstanding scientists were defined as those who had published $N$ or more papers – a criteria put forward by Price. Formally, the calculation is $N = 0.749(\beta_{max})^{1/2}$, where $\beta_{max}$ is the highest number of papers published by any author in the dataset. Thus, we selected 813 researchers with five or more publications for future analysis.

From sections 3.2 to 3.4, we selected data from 2010 to 2013 to complete the remaining three main steps of the recommendation framework. To further verify the quality of recommendations, from Section 4.5, we divided the dataset into records from 2010 to 2013 as the training set and 2014 to 2018 as the testing set, and make a comparison with the two most commonly used baselines: RWR and LDA.

## 4.2 Constructing the network of research interests with Word2Vec

In this section, first, the parameters of the Word2Vec model were set to a window size of 2 and a layer size of 128 based on the testing of a number of options. We then generated the research interest vectors for all scholars according to Eq. (1) and Eq. (2). Eq. (3) subsequently gave us an 813 × 813 symmetrical similarity matrix of research

interests. Alexey Matveev and Andrey Savkin had the most similar interests at 0.981981, Senqiang Zhu and Frederic Py had the least similar at 0.002712. The mean value, median and standard deviation values were 0.544555, 0.580275 and 0.209891, respectively.

Fig. 2 shows the network scholars with a similarity score of more than 0.55. The size of the node represents the number of published papers for each scholar, and the thickness of the lines indicates the degree of similarity between the scholars' research interests.

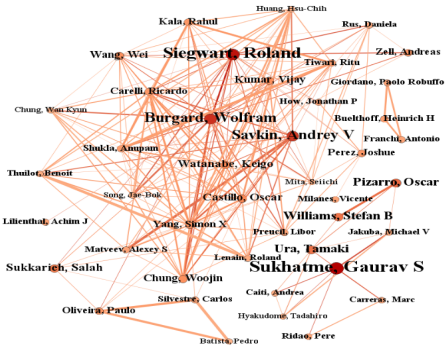


**Fig. 2.** The similarity network of research interests

### 4.3 Constructing the co-authorship network with Node2Vec

In this section, first, the parameter settings for the Node2Vec model were also determined from testing. The final values were: dimensions=128; walk-length=80; p=1; and q=1. Eq. (3) yielded the 813 × 813 topology matrix of cosine similarity between scholars, which was again symmetrical. Gaurav S Sukhatme and Ryan N Smith shared the greatest similarity (0.957244), and Jian Liu and Yi Chao had the least (0.028196). The mean, median and standard deviation values were 0.416515, 0.414858, and 0.125672, respectively.

Fig. 3 shows the co-authorship network based on scholars with a topological similarity of more than 0.47. The size of nodes indicates the number of collaborators associated with that scholar. The thickness of the lines represents the degree of similarity between the two connected scholars.

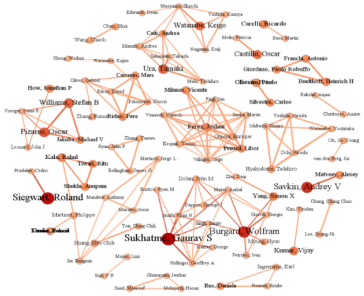

**Fig. 3.** The co-authorship network showing the topological similarities between scholars

### 4.4 Ranking the candidate collaborators with CombMNZ

As a preliminary assessment of the framework's ability to make appropriate recommendations, we randomly selected Roland Siegwart as the target scholar and generated a final list of recommendations. The top-10 ranked candidates are shown in Table 1.

**Table 1.** The top 10 recommended collaborators for Roland Siegwart based on the 2010-2013 dataset

| No | Recommended Collaborator | No | Recommended Collaborator |
|---|---|---|---|
| 1 | Weiss, Stephan | 6 | Posner, Ingmar |
| 2 | Stachniss, Cyrill | 7 | Bedkowski, Janusz |
| 3 | Newman, Paul | 8 | Scaramuzza, Davide |
| 4 | Liu, Ming | 9 | Mondada, Francesco |
| 5 | Pradalier, Cedric | 10 | Gonzalez, Ramon |

An in-depth manual review of Siegwart's academic background shows these recommendations to be appropriate. For example, Stachniss and Siegwart have overlapping interests in mobile robots, sensor design, navigation system design, positioning, motion planning and more, and have both published many influential papers. In addition, both scholars often attend the IEEE International Conference on Robotics & Automation. Based on this analysis, it is reasonable to conclude that the framework can recommend realistic and fruitful collaborations.

### 4.5 Comparative evaluations

From the literature, we found the two most popular and widely-used methods for collaborator recommendations are the LDA model and the RWR indicator coupled with a topological model. To compare the quality of recommendations produced by these approaches with those of our framework, we divided the dataset into records from 2010 to 2013 as the training set and 2014 to 2018 as the testing set. And then we randomly chose 20 target authors for comparison by Precision, Recall, and F1 scores. The results are given in Figs. 4 to 6.

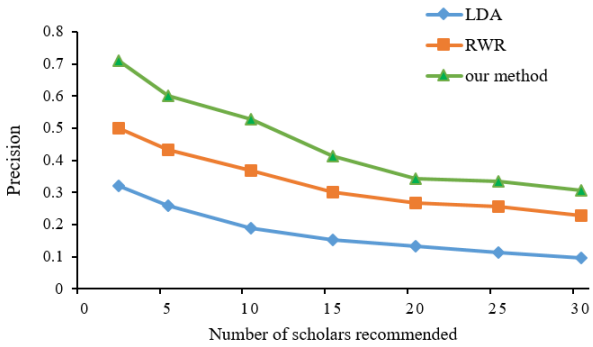


**Fig. 4.** Precision

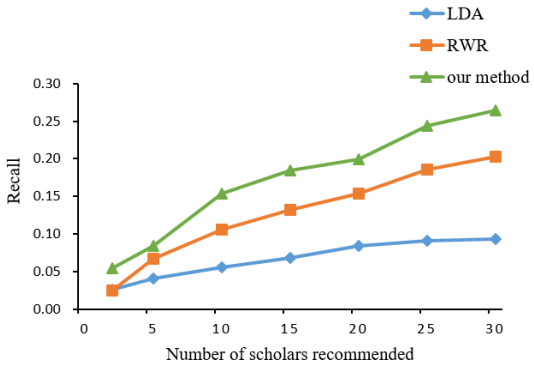


**Fig. 5.** Recall

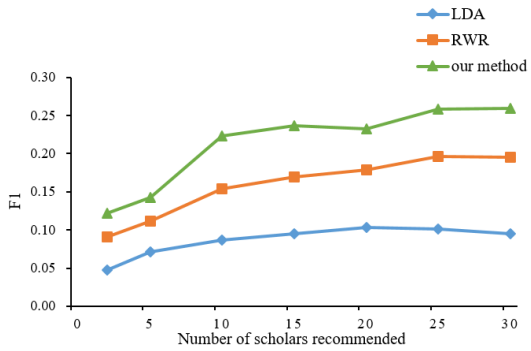


**Fig. 6.** F1 Score

In Figs. 4 to 6, we find that our framework makes higher precision, recall and f1 score than the two current benchmark solutions, these results emphasize the advantages of our approach. More specifically, we drew the following insights from this analysis:

(1) Fusing similarity indicators based on research interests and topological structures significantly increased the quality of the recommendations.

(2) The Word2Vec method solved the problems of context less and scalability associated with traditional text mining technology.

(3) The Node2Vec method removed the need to manually design and define indicators, saving on manpower and produced recommendations based on global network features.

## 5 CONCLUSION

Overall, the main innovation of our paper is to develop a novel framework for recommending academic collaborators with similar research interests and network topology features through incorporating word embedding and network embedding, and produces more accurate recommendation compared with the existing methods. Meanwhile, the framework can not only help researchers and private-enterprise R&D managers to provide valuable reference for cooperation, but also is feasible and can now be the basis for further improvement/inspiration.

The limitations of our current research offer opportunities for future inquiry. These are summarized as follows. (1) Word embedding and network embedding techniques both contain some parameters; however, methods of training these parameters for optimal benefit is a task that falls into the field of machine learning. (2) We have based our recommendations on only two criteria: the similarity of research interests and the co-authorship features. However, other factors can also indicate the likelihood of a good collaboration, such as citations, or institutional ties. In future, we will consider adding more of these factors into our framework.

## References

1. Katz, J. Sylvan, & Martin, Ben R.: What is research collaboration?. Research Policy 26(1), 1-18 (1997).
2. Sooho, L., & Barry, B.: Scientific collaboration || the impact of research collaboration on scientific productivity. Social Studies of Science 35(5), 673-702 (2005).
3. Guns, R., & Rousseau, R.: Recommending research collaborations using link prediction and random forest classifiers. Scientometrics 101, 1461-1473 (2014).
4. Abramo, G., D'Angelo, C. A., & Costa, F. D.: Research collaboration and productivity: is there correlation? Higher Education 57(2), 155-171 (2009).
5. Kong, X., Jiang, H., Wang W. et al.: Exploring dynamic research interest and academic influence for scientific collaborator recommendation. Scientometrics 113(1), 369-385 (2017).
6. Xia, F., Chen, Z., Wang, W., Li, J., & Yang, L. T.: Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors. IEEE Transactions on Emerging Topics in Computing 2(3), 364-375 (2014).
7. Pham, M. C., Cao, Y., Klamma, R., & Jarke, M.: A clustering approach for collaborative filtering recommendation using social network analysis. Journal of Universal Computer Science 17(4), 583-604 (2011).

8. Dong, Y., Tang, J., Wu, S., et al.: Link Prediction and Recommendation across Heterogeneous Social Networks. 2012 IEEE 12th International Conference on Data Mining. IEEE. pp. 181-190 (2013).
9. Ping, N., & De-Gen, H. Tf-idf and rules based automatic extraction of chinese keywords. Journal of chinese computer systems (2016).
10. Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5), 993-1022 (2008).
11. Le, Q., & Mikolov, T. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning (ICML-14), 1188-1196 (2014).
12. Zhang, Q., Xu, X., Zhu, Y., & Zhou, T. Measuring multiple evolution mechanisms of complex networks. Scientific Reports, 5, 10350 (2015).
13. Kong, X., Jiang, H., Wang W. et al. Exploring dynamic research interest and academic influence for scientific collaborator recommendation. Scientometrics 113(1), 369-385 (2017).
14. Peng, C., Xiao, W., Jian, P., & Wenwu, Z. (2018). A survey on network embedding. IEEE Transactions on Knowledge and Data Engineering, pp (2018).
15. Zhang, Y., Porter, A.L., Hu, Z., Guo, Y., & Newman, N.C.: "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells. Technol. Forecast. Soc. Chang 85, 26-39 (2014a).
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 3111-3119 (2013).
17. Eunice T., Iris S., Humphrey L., & Yiu-Kai N. Making personalized movie recommendations for children. International Conference on Information Integration & Web-based Applications & Services. ACM (2016).
18. Macdonald, C., Ounis, I.: Voting techniques for expert search. Knowledge & Information Systems, 16(3): 259-280 (2008).
19. Kwon, S., Liu X., Porter, A.L., Youtie, J.: Research addressing emerging ideas has greater scientific impact. Research Policy 48(9), 1-1 (2019).